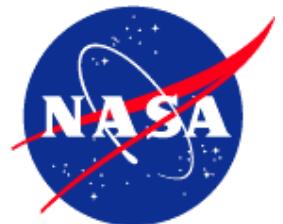


Machine-Aided Indexing

Bill von Ofenheim

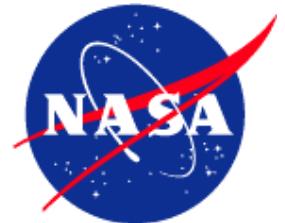
NASA Langley Research Center

w.h.c.vonofenheim@larc.nasa.gov



Overview

- MAI Definition
- Example
- Features
- Algorithm
- NASA Thesaurus
- Knowledge Base
- Implementation
- Future



MAI Definition

- Computer-assisted method to derive index terms (or keywords) from a technical document
- Index terms are standardized on the NASA thesaurus
- URL: <http://mai.larc.nasa.gov/>



Example

Netscape: NASA Thesaurus Machine Aided Indexing

Bookmarks Location: <http://mai.larc.nasa.gov/>

NASA Scientific and Technical Information (STI) Program
NASA Thesaurus Machine Aided Indexing (MAI) STI

Perform Indexing Clear Form [Instructions](#) | [Home](#)

The present paper studies the numerical simulation of flows with shock/boundary-layer upstream interaction, under conditions of symmetry in geometry, boundary conditions, and grid. For this purpose, a series of two- and three-dimensional numerical test-cases were carried out. The tests showed that standard numerical schemes, which appear to be symmetry-preserving under most flow

Freq.	NASA Term
3	symmetry save
2	perturbation save
1	shock wave interaction save
1	boundary layers save
1	upstream save
1	boundary conditions save
1	core flow save
1	separated flow save

Enter word: [Search](#) [Instructions](#)

Thesaurus Hierarchical Display

TERM [boundary conditions](#) [remove](#)

HIERARCHY

[conditions](#) [save](#)

. [boundary conditions](#) [remove](#)

. . [perfectly matched layers](#) [save](#)

RELATED TERMS

[boundaries](#) [save](#)

[boundary element method](#) [save](#)

[boundary layers](#) [save](#)

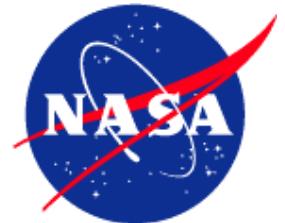
[boundary value problems](#) [save](#)

[Treffitz method](#) [save](#)

[vortex lattice method](#) [save](#)

Saved Terms (Add terms by clicking on [save](#)) [clear](#)

[boundary conditions](#) ; [shock wave interaction](#) ;



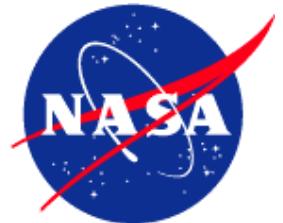
Features

- Submit full-text documents
- Ordered frequency of occurrence counts
- Lists unrecognized words
- Single word search of NASA thesaurus
- Navigation through thesaurus
- Shopping-cart capability to save preferred terms



Algorithm

- Clean submitted text and break into punctuation and stop word delimited sub-sentences
- Use sliding, variable window to process
- Match to phrases in Knowledge Base
- Mark matched words as “poisoned” to prevent duplicate usage
- Map to NASA thesaurus terms and generate frequency counts



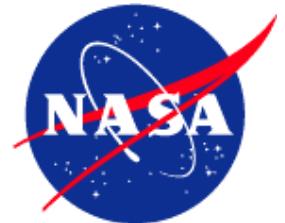
NASA Thesaurus

- NASA's controlled vocabulary of ~18K terms
- Arranged in hierarchical fashion
- Cross-referenced preferred usages
- Includes related terms
- Limited definitions



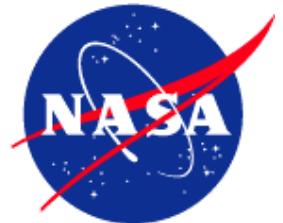
Knowledge Base

- Corpus analysis to cull candidate phrases (~150K)
- Phrases connected to associated thesaurus terms
- Example:
 - Term: **aeroacoustics**
 - acoustically suppressed exhaust nozzle
 - aero acoustic
 - aero-acoustic
 - aerodynamic sound
 - edge-tone effect
 - flap noise
 - propeller tone burst
 - turbofan engine noise radiation



Implementation

- Created as Java servlets
 - Provides persistence, session management, and device independence
- 5 Databases (Thesaurus Volume 1 and 2, Thesaurus Index, Knowledge Base, and Stop Words) loaded into core
- Data structures optimized for fast access while minimizing storage costs
- Separate Java client for batch processing



Future

- Include subject category determination
- Integrate with Report Documentation Page (RDP)